Interdisciplinary spatial downscaler modeling approach:
Enhancing air quality measurement predictions in central California

Tom Gause, Zahra Ameli, Heather Clifford

1.Introduction

1.1 Problem Statement

Lack of ground-based air quality monitor systems creates discontinuity for spatially extensive measurements, leading to gaps in our understanding of air quality levels in different areas. We apply a static version of a downscaler model using the combination of air quality measurements from a chemical transport model and monitoring stations in order to yield improved spatial predictions of air pollutant levels at a more local scale.

1.2 Motivation

A rapidly changing climate has caused an increase in frequency of wildfires in the western United States, in addition to longer durations and a greater area affected (Marlon et al., 2012; Westerling et al., 2006). Wildfires are a major source of ash, also known as particulate matter less than 2.5 **u**m (PM2.5), which causes diminished air quality to local areas and is only projected to get worse in time (Jaffe et al., 2008). PM2.5 can cause adverse health effects through chronic, low level exposure, some health impacts include cardiovascular and respiratory morbidity and mortality (Brook et al., 2010; Pope & Dockery, 2006).

Air quality (e.g. PM2.5) is measured using several different methods including individual ground-based monitoring stations (located in areas of high population density), chemical transport models (CTMs), and satellite-based aerosol optical depth (AOD). While ground-based monitoring stations are beneficial to assess accurate PM2.5 measurements at specified locations, they are spatially sparse, causing lack of accurate high-resolution spatial measurements. Another method used to measure daily surface PM2.5, which has a higher spatial resolution, is the Community Multiscale Air Quality model (CMAQ), created by the National Oceanic and Atmospheric Administration (NOAA) and based on a meteorological model of North America. Using spatial-temporal models to predict PM2.5 levels can help to decrease exposure misclassification and allow for more accurate assessments of health exposure risks. Additionally, using combinations of existing measurements and models can further the precision of such models.

Our **motivation** for this project is to create an interdisciplinary modelling approach by combining environmental datasets and data science methodology to enhance the precision of air quality estimations, which in turn can help alert the general public of increased risk related to poor air quality. We base our investigation for this project from several studies that predict PM2.5 measurements spatially and temporally by using a Bayesian Ensemble model to combine the datasets from monitoring stations, CMAQ model outputs and satellite-based AOD (Geng et al., 2018, Murray et al., 2019). Due to the magnitude of work required for such a model, we reduced our model approach to a static downscaler method, which also uses a combination of

ground-based monitoring stations and CMAQ model outputs. We chose central California (35.0 to 38.2 N, -121.7 to -117.8 W) as the spatial extent of our study due to the magnitude of recent wildfires occurring in the area, high population density and lack of spatial monitoring systems, shown in **Figure 1**. The temporal extent of our study is the average daily PM2.5 measurements for January 1, February 14, and December 30, 2016.

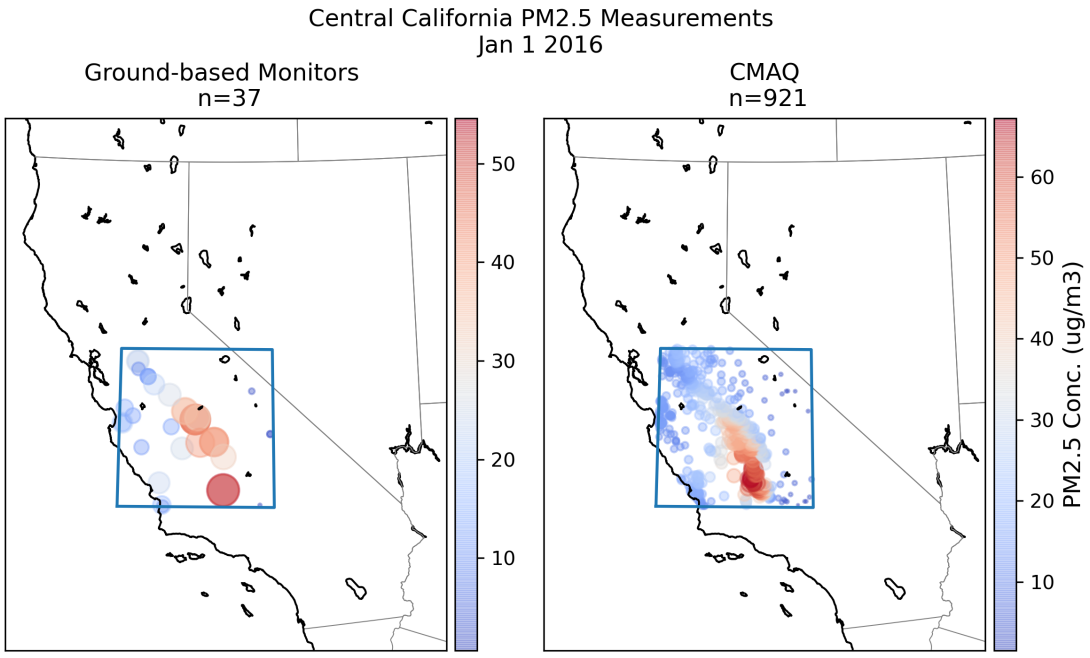

Figure 1. PM2.5 measurements for Central California on January 1, 2016 from ground-based air quality monitoring systems and CMAQ model outputs.

1.3 Related Work

The main idea for this project was extracted from Geng et al. (2018), which utilizes a Bayesian Ensemble Model (BEM) to predict $PM_{2.5}$ concentrations during fire season in Colorado from 2011-2014. This paper used daily 24-hr mean $PM_{2.5}$ concentration data at 46 monitoring stations from the U.S. EPA's Air Quality System to estimate concentrations at 1 km resolution with reasonable success. Similarly, Murray et al. (2019) combines statistical data fusion models using either chemical transport model (CTM) simulation or satellite AOD through Bayesian ensemble averaging to predict daily PM2.5 in the Southeastern United States. We deviate from the BEM approach by building a similar but simplified static model using the same databases (excluding AOD measurements) at a different location and a specified time window. To assist in building the model, we apply an adjusted methodology guided by Berrocal et al. (2010), which uses spatio-temporal Bayesian downscaler framework on similar datasets to produce predicted ozone measurements. Due to the obstacles arising from interdisciplinary work, we find Berrocal et al. (2010) describes the methodological approach in more significant detail, comparatively.

Although regression based models can provide reasonable results, the non-linear behavior of air pollutants and other influential regional features leads to a very complex system of air pollutant formation. For that, advanced statistical approaches based on machine learning algorithms (e.g. neural network (NN), support vector machine (SVM) and Ensemble Learning algorithms) are developed which yield better results. Due to wide applications in air quality modeling, NNs are considered one of the most common, reliable, widely adopted and cost effective machine learning tools to predict air pollutant concentrations (Russo and Soares, 2014; Shaban et al., 2016; Rahimi, 2017). Some recent works have proposed hybrid models and have claimed their robust performances in severe pollution scenarios (Tamas et al., 2016; Perez and Menares, 2018). For example in a recent study by Karimian et al. (2019), the LSTM model which is a hybrid model based on long short-term memory was able to capture temporal dependencies in time series data, which increased the accuracy of its PM2.5 forecasting.

2. Methodology

2.1 Data

Daily 24-hr mean PM2.5 concentration data from 2016-2017 for Central California was downloaded from the U.S. Environmental Protection Agency's Air Quality System (www.epa.gov/outdoor-air-quality-data/). For the selected location (35.0 to 38.2 N, -121.7 to -117.8 W), there are 45 PM2.5 monitoring stations, shown in **Figure 1a,** that met pre-processing guidelines of >300 daily measurements for the years 2016 and 2017. We additionally downloaded CTM simulations for daily PM2.5 (24-hr average) from 2016-2017 from the USEPA Models-3/CMAQ model (www.epa.gov/hesc/rsig-related-downloadable-data-files) at a 12 km × 12 km horizontal spatial resolution (Byun and Schere, 2006). For central California, there are 916 measurements available for both 2016 and 2017, shown in **Figure 1b**. Once downloaded, the datasets were pre-processed and set in a gridded matrix form for model input using python.

2.2 Static Spatial Downscaler Model

For this study, we apply a static spatial downscaler model, specified within a Bayesian framework, to downscale the CMAQ numerical model output to point level, defined by Berrocal et al. (2010). This spatially-varying weighted model regresses observed monitoring data on the CMAQ output using spatially varying coefficients specified through a correlated spatial Gaussian process. This process takes advantage of the CMAQ spatial information in the gridded cells that surround where the monitor data is collected and helps to create a more reliable association between the two. Our decision to use a static (fixed condition) model is to simplify the methodologies proposed in both Geng et al., (2018) and Murray et al., (2019), due to the complexity of using spatially and temporally expansive datasets, in addition to removing the numerous obstacles that are involved in Bayesian Ensemble Models.

Here, we define the steps used to build the static downscaler model, derived from the methodology in Berrocal et al (2010). We denote the square root of the observed PM2.5 concentration at a point **s** with $Y(\mathbf{s})$. Each point **s** is associated with the 12x12 km CMAQ grid cell B in which it lies. Since CMAQ output is given in terms of averages over 12x12 km grid cells, we use $x(B)$ to denote the square root of the numerical model output over grid cell B. Therefore, all the point**s** falling in the same 12 km square region are assigned the same CMAQ output value. We relate the observed data to the CMAQ output in the following way: for each **s** in B, we assume that:

$$Y(s) = \tilde{\beta}_0(s) + \tilde{\beta}_1(s)x(\beta) + \varepsilon(s), \quad \varepsilon(s) \sim N(0, \tau^2)$$

Where:

$$\tilde{\beta}_0(s) = \beta_0 + \beta_0(s),$$
$$\tilde{\beta}_1(s) = \beta_1 + \beta_1(s),$$

$\varepsilon(s)$ is a white noise process with nugget variance τ^2. β_0 and β_1 represent the overall additive and multiplicative bias of the CMAQ model, while $\beta_0(s)$ and $\beta_1(s)$ are local adjustments to the additive and multiplicative bias, respectively. We suppose that there exist two mean-zero unit-variance independent Gaussian processes $w_0(s)$ and $w_1(s)$ such that, for convenience, $\text{cov}(w_j(s), w_j(s')) = \exp(-\phi_j |s - s'|)$, ϕ_j is the spatial decay parameter for Gaussian process $w_j(s)$, $j = 0, 1$, i.e.

$$cov(w_j(s), w_j(s')) = \begin{pmatrix} |s_0 - s_0| & |s_0 - s_1| & \dots & |s_0 - s_n| \\ |s_1 - s_0| & \dots & \dots & \dots \\ & \dots & \dots & \dots \\ |s_n - s_0| & \dots & \dots & |s_n - s_n| \end{pmatrix}$$

for n sites and:

$$\begin{pmatrix} \beta_0(s) \\ \beta_1(s) \end{pmatrix} = A \begin{pmatrix} w_0(s) \\ w_1(s) \end{pmatrix}$$

The unknown **A** matrix can be assumed, without loss of generality, to be a lower triangular. This model specification is rather simple yet it provides calibration at the local level and endows the spatial process Y(s) with a flexible nonstationary covariance structure. This model is much easier to fit than Bayesian melding since we eliminate the need to evaluate stochastic integrals. Moreover, to fit the model we only need to work with the responses associated with the Y(si), i.e., with the set of monitoring sites, a relatively small number compared to the number of grid cells. We evaluated our downscaling model by looking at the mean squared error (MSE), mean absolute error (MAE), and variance of the predictions at the validation sites.

3. Results & Discussion

We chose the topic of this paper before we understood the complexity of bayesian atmospheric modeling. Originally, we chose a Bayesian hierarchical model combining AOD, CMAQ, and monitoring station data to make spatio-temporal predictions, thinking we could easily take a model from another paper and interpolate it to our new locations and times. We quickly discovered that not only are these models extremely complicated to brand new data science students, but these interdisciplinary papers not only did not include any actual code, they barely explained the models used. Thus, the more we learned, the more we simplified our project. In the end, we chose to implement a Bayesian static downscaler model, one of the models in the BHM pipeline. This model can be scaled to a spatio-temporal model by including additional Gaussian processes and training additional additive/multiplicative parameters. This spatio-temporal model would have been fairly easy to implement, but we lack the necessary computation resources to run the model. For example, generating a single sample from our multivariate normal distribution at 2k resolution took over 90 minutes on our machines. To scale to time effectively, we would need to generate several thousand samples for each day - this was not possible.

We took two further steps to simplify our model. First, we assumed that the multiplicative bias of the numerical model output is constant in space, thus we set the local adjustment $\beta_1(s)$ to zero. Second and more significantly, we fixed *all* parameters in the model: β_0, β_1, A_{11}, ϕ_j, and τ^2. In the Berrocal paper, these parameters were estimated via MCMC with a combination of Gibbs and Metropolis-Hastings steps. Instead of the MCMC, we took the mean of their parameters after 10,000 iterations as our own fixed priors ($\beta_0 = 3.2$, $\beta_1 = 0.5$, $A_{11} = 1.1$, $\phi_0 = 0.003$, and $\tau^2 = 0.1$). This was a necessary step, as we lack the computational resources to implement the MCMC ourselves - each step would require an additional sample from the Gaussian processes and accurate results take thousands of iterations, therefore several months of compute time. We decided against manually tuning these parameters for improved performance as working with five degrees of freedom is a far more machine-oriented task.

As a result of these modifications, our model generated a fairly accurate high resolution dataset with local variety introduced by the Bayesian process (**Figures 2 and 3**). We did not expect to get significant or accurate results. The MSE for 2016 (**Table 1**) indicates our projections were correct - the CMAQ values, which the objective of the Berrocal downscaler was to improve upon, fits the monitoring station data far better than our downscaler does. We do not consider this a failure.

Table 1. Parameters and performance evaluation for each day chosen to run the model.

Day	1/1/2016	2/14/2016	12/30/2016
Method	Downscaler	Downscaler	Downscaler
# of CMAQ Measurements	921	921	921
# of Monitoring Systems (MS)	37	29	31
Size of Model Output (MO)	7921	7921	7921

Resolution of Output	4km x 4km	4km x 4km	4km x 4km
Avg. MSE (MO and MS)	82.71	93.68	145.94
Avg. MSE (CMAQ and MS)	14.16	6.24	35.7
Avg. MSE (CMAQ and MO)	75.90	83.56	97.46
Avg. MAE (MO and MS)	8.13	8.59	11.06
Avg. MAE (CMAQ and MS)	2.75	1.69	4.07
Avg. MAE (CMAQ and MO)	8.03	8.34	8.86

4. Conclusions

We find the spatial static downscaler model approach to be beneficial for the purpose of improving spatial predictions of air pollutant levels at a more local scale. This application is not without difficulty due to the interdisciplinary nature of the approach and numerous moving parameters that are typically included within the model are not fixed. However, we were able to successfully apply our model to the selected datasets from the ground-based monitoring systems and CMAQ model outputs and achieve our goal of expanding the spatial resolution of our dataset (**Figures 2 and 3**). The performance of our model output (**Table 1**) is significantly worse than the results from Berrocal et al. (2010); however, these differing results are not unexpected since our model is a simplified version with fixed parameters. Future improvements in the choice of indicators and accuracy of input data will likely further enhance the model's performance. For example, the PM2.5 simulations from the CMAQ model in our study domain were underestimated compared to the observation data, which was caused by the uncertainties in the model emissions and meteorological inputs. Improvements in the CMAQ model are very likely to improve the performance of the CMAQ downscaler and reduce the model biases in the final predictions.

5. Group member roles

Zahra's role is to find references relevant to the project, working on the data and Bayesian ensemble model and work on the project paper. Tom's role is to research and implement methodology, manage/augment data, build the model, and work on the project paper. Heather's role for the group project is to download and preprocess datasets, write introduction and assist with methodology, contribute to coding for data visualization, and add relevant references.

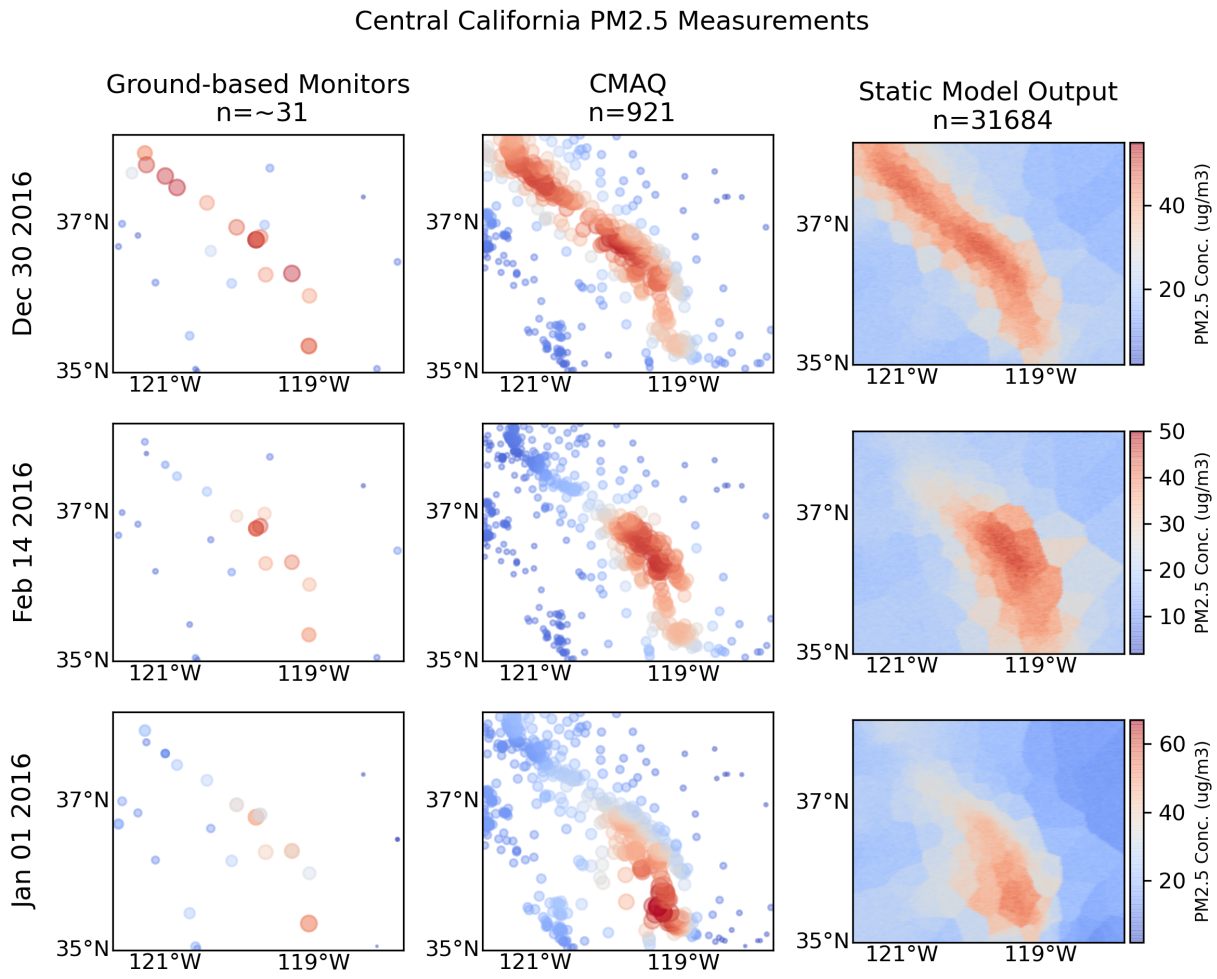

Figure 2. Overview of PM2.5 datasets from our study for Central California on December 30, February 14 and January 1 of 2016, including the ground-based air quality monitors, CMAQ model outputs and the static downscaler model output results with 2km by 2km resolution.

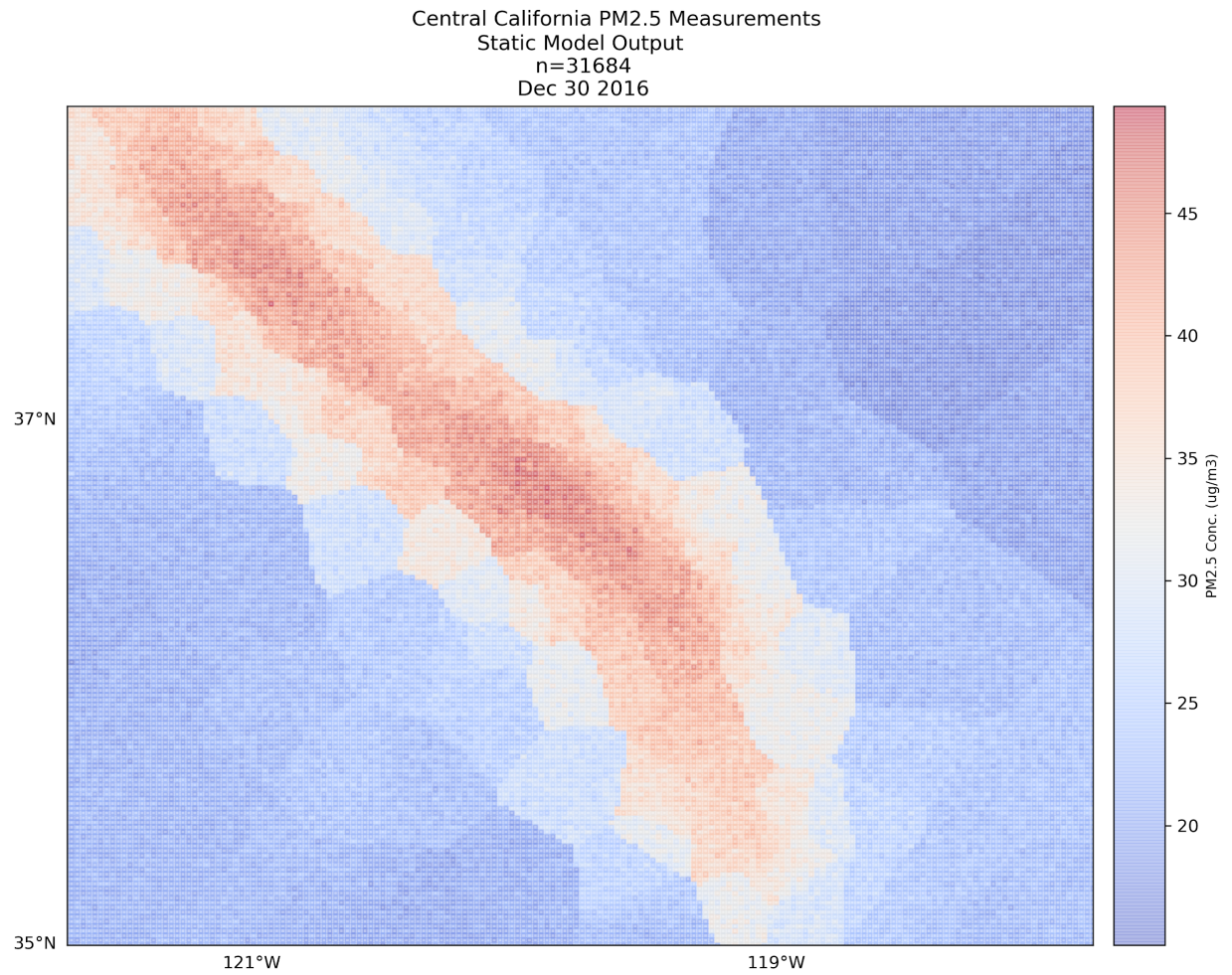

Figure 3. Focused view of static downscaler model output results with 2km by 2km resolution for December 30, 2016.

References

1. Masih A. Machine learning algorithms in air quality modeling. *Global Journal of Environmental Science and Management* 2019; **5**(4). DOI: 10.22034/GJESM.2019.04.10.
2. Mirzaei M, Bertazzon S, Couloigner I, Farjad B, Ngom R. Estimation of local daily PM2.5 concentration during wildfire episodes: integrating MODIS AOD with multivariate linear mixed effect (LME) models. *Air Quality, Atmosphere & Health* 2020; **13**(2): 173–185. DOI: 10.1007/s11869-019-00780-y.
3. Geng G, Murray NL, Tong D, Fu JS, Hu X, Lee P, *et al.* Satellite-Based Daily PM $_{2.5}$ Estimates During Fire Seasons in Colorado. *Journal of Geophysical Research: Atmospheres* 2018. DOI: 10.1029/2018JD028573.
4. Westerling AL, Hidalgo HG, Cayan DR, Swetnam TW. Warming and Earlier Spring Increase Western U.S. Forest Wildfire Activity 2006; **313**: 4.

5. Pope CA, Dockery DW. Health Effects of Fine Particulate Air Pollution: Lines that Connect. *Journal of the Air & Waste Management Association* 2006; **56**(6): 709–742. DOI: 10.1080/10473289.2006.10464485.
6. Murray NL, Holmes HA, Liu Y, Chang HH. A Bayesian ensemble approach to combine PM2.5 estimates from statistical models using satellite imagery and numerical model simulation 2020: 20.
7. Murray N, Chang HH, Holmes H, Liu Y. Combining Satellite Imagery and Numerical Model Simulation to Estimate Ambient Air Pollution: An Ensemble Averaging Approach. *ArXiv:180203077 [Stat]* 2018.
8. Marlon JR, Bartlein PJ, Gavin DG, Long CJ, Anderson RS, Briles CE, *et al.* Long-term perspective on wildfires in the western USA. *Proceedings of the National Academy of Sciences* 2012; **109**(9): E535–E543. DOI: 10.1073/pnas.1112839109.
9. Karimian H, Li Q, Wu C, Qi Y, Mo Y, Chen G, *et al.* Evaluation of Different Machine Learning Approaches to Forecasting PM2.5 Mass Concentrations. *Aerosol and Air Quality Research* 2019; **19**(6): 1400–1410. DOI: 10.4209/aaqr.2018.12.0450.
10. Just AC, Wright RO, Schwartz J, Coull BA, Baccarelli AA, Tellez-Rojo MM, *et al.* Using High-Resolution Satellite Aerosol Optical Depth To Estimate Daily PM $_{2.5}$ Geographical Distribution in Mexico City. *Environmental Science & Technology* 2015; **49**(14): 8576–8584. DOI: 10.1021/acs.est.5b00859.
11. Jaffe D, Hafner W, Chand D, Westerling A, Spracklen D. Interannual Variations in PM2.5 due to Wildfires in the Western United States. *Environmental Science & Technology* 2008; **42**(8): 2812–2818. DOI: 10.1021/es702755v.
12. Deters JK, Zalakeviciute R, Gonzalez M, Rybarczyk Y. Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. *Journal of Electrical and Computer Engineering*: 15.
13. Brook RD, Rajagopalan S, Pope CA, Brook JR, Bhatnagar A, Diez-Roux AV, *et al.* Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement From the American Heart Association. *Circulation* 2010; **121**(21): 2331–2378. DOI: 10.1161/CIR.0b013e3181dbece1.
14. Berrocal VJ, Gelfand AE, Holland DM. A Spatio-Temporal Downscaler for Output From Numerical Models. *Journal of Agricultural, Biological, and Environmental Statistics* 2010; **15**(2): 176–197. DOI: 10.1007/s13253-009-0004-z.
15. Russo, A., Soares, A.O. Hybrid Model for Urban Air Pollution Forecasting: A Stochastic Spatio-Temporal Approach. Math Geosci 46, 75–93 (2014). https://doi.org/10.1007/s11004-013-9483-0
16. Sayegh, A.S., Munir, S. and Habeebullah, T.M. (2014). Comparing the Performance of Statistical Models for Predicting PM10 Concentrations. Aerosol Air Qual. Res. 14: 653-665. https://doi.org/10.4209/aaqr.2013.07.0259
17. Rahimi, A. Short-term prediction of NO2 and NO x concentrations using multilayer perceptron neural network: a case study of Tabriz, Iran. Ecol Process 6, 4 (2017).

https://doi.org/10.1186/s13717-016-0069-x
18. Tamas, W., Notton, G., Paoli, C., Marie-Laure Nivet, M.L. and Voyant, C. (2016). Hybridization of air quality forecasting models using machine learning and clustering: An original approach to detect pollutant peaks. Aerosol Air Qual. Res. 16: 405–416.
19. Perez, P. and Menares, C. (2018). Forecasting of hourly PM2.5 in south-west zone in Santiago de Chile. Aerosol Air Qual. Res. 18: 2666–2679.